DOCUMENT RESUME

ED 170 363 TM 008 821

AUTHOR Kelley, Paul R.
TITLE Combined Common Person and Common Item Equating of Medical Science Examinations.
PUB DATE 11 Apr 79
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979) ; Figure 1 may be marginally legible due to type

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Complexity Level; *Equated Scores; *Family Characteristics; Goodness of Fit; Longitudinal Studies; *Measurement Techniques; Occupational Surveys; Student Ability; Test Items
IDENTIFIERS *National Board Examinations; Rasch Model

ABSTRACT

This equating study of the National Board of Medical Examiners Examinations was a combined common persons and common items equating, using the Rasch model. The 1,000-item test was administered to about 3,000 second-year medical students in seven equal-length subtests: anatomy, physiology, biochemistry, pathology, microbiology, pharmacology, and behavioral science. Scores were equated through common items to those on a similar reference examination administered previously to approximately 8,500 second-year students. Although scores are reported for each of the seven subtests, pass or fail decisions toward certification are based only on the total score. Items on the previously administered reference subtests were calibrated and equated by shifting the item difficulties to a common scale such that each test would produce the same estimate of ability for the sample tested, resulting in a mean ability score of zero for the examinees. Current test items were also calibrated; link analysis indicated differences in current and reference test items. Item difficulties were then adjusted to equate the current examination to the reference test. (A second equating study is also described). (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED170363

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

COMBINED COMMON PERSON AND COMMON ITEM EQUATING OF MEDICAL SCIENCE EXAMINATIONS

Paul R. Kelley
National Board of Medical Examiners

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Paul R. Kelley

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Presented at the 1979 AERA/NCME
April 11, 1979

SESSION #22.20

TM008 821

COMBINED COMMON PERSON AND COMMON ITEM EQUATING OF MEDICAL SCIENCE EXAMINATIONS

The first major equating study of National Board Examinations using the Rasch model was a combined common persons and common items equating. The study was done on an official administration of a 1,000 item national examination in the basic medical sciences administered to about three thousand second-year medical students. This total examination was made up of seven equal length subtests representing the basic science disciplines generally taught in the first one or two years of medical school. Scores on this test, which I shall refer to as the current test, were to be equated through common items to those on a similar reference examination administered previously to approximately 8,500 second-year medical students. Scores are reported for each of the seven subtests but pass or fail for purposes of credit toward certification for a license to practice medicine is based only on the total test score.

The first step in the equating procedure was to calibrate the items on the previously administered reference test. This was done for each of the seven discipline subtests separately. The mean person abilities in logits for each subtest are shown as the first column of figures in Table 1. Since the same persons took all seven subtests and since all subtests measure the same underlying variable, namely, knowledge in the basic medical sciences, the differences in mean ability by subtest must be due to differences in the difficulty of the items constituting each subtest. Therefore, one could equate these subtests simply by shifting the item difficulties to a common scale such that each test would produce the same estimate of ability for the sample tested. For example, the Anatomy item difficulties would be reduced by .51 logits, the Physiology items by .79 logits, etc. The effect of this common person equating would be

to produce a mean ability of zero for this sample of examinees on each of the seven subtests.

So much for the common person equating of several subtests or components of a single examination. Referring again to the equating of the 1,000 item current examination and each of its subtests to the reference test, the next step, having calibrated the items of the reference test, was to calibrate the items of the current examination. Once the reference test and current test average item diffiuclties for the common items are in hand, a link analysis determines the difference between them after the exclusion of a few items that fit poorly with the link. Since the items are the same, this difference has to be due to the difference in the two arbitrarily established origins of the item difficulty scales for the reference and current examinations. Adjustment of all the current examination item difficulties by the amount of the difference would equate them to the scale of the reference examination. These differences by subtest may be found in the second column of Table 1. These figures were arrived at by subtracting the mean common item difficulties of the current test from those of the reference test. Thus, the three subjects with the negative differences are ones in which all items, on the average, were easier in the current subtest than in the corresponding reference subtest. For Anatomy, there was essentially no difference. For the remaining three subtests, the items on the current test were more difficult than those on the reference test.

The third column in Table 1 is the difference between the values in columns 1 and 2 and is the shift in the current test item difficulties needed to equate those difficulties to the common origin of the reference examination.

Thus, this shift incorporates both the common person equating used to place the item difficulties of all seven subtests at a single origin and the common item equating that adjusts item difficulties on the current examination to those of the reference test. As may be seen in Table 1, it is the first column of figures, i.e., the adjustment needed to equate the subtests within the reference examination, that is the primary determinant for equating the current examination to a common origin on the reference examination scale.

The common person equating in this particular combined common person, common item procedure was necessary only to facilitate the calibration of a 1,000 item test. Incidentally, the equating procedure just described was carried out at the University of Chicago using BICAL and other programs developed by Benjamin Wright and Ronald Mead. For the second major equating study on National Board examinations, the calibration program was modified by Ronald Mead and David Kreines to enable it to calibrate all 1,000 items in a single calibration run, placing the item difficulties at a common origin without the necessity of calibrating each subtest separately and then using the common person procedure to equate the subtests. All analyses involved in this second equating study were done on the National Board computer in Philadelphia.

The equating problem for this second study was essentially the same as the one just described. A 1,000 item examination in the basic medical sciences and its several subtests was to be equated to a previously administered reference examination. The entire reference test was calibrated as a single examination. The same procedure was followed in calibrating the current examination. Therefore, the difficulties of all items within each of the two examinations were calibrated to a common origin. The equating of the current examination to the reference test, the entire examination or any component of it, was done through

the use of one shift constant determined by the link analysis on the items common to both tests. So, rather than all the figures shown in Table 1 being involved in the equating, there was only a single number. All item difficulties on the current examination were shifted by .58 logits to move them to the single origin of the reference scale.

Use of the Rasch model assumes that the test content to which it is being applied meets statistical criteria for unidimensionality. Before any equating studies using the model were done on National Board examinations, the question of how well the content of the tests fit the model was addressed. Thousands of items from basic and clinical science examinations were calibrated on a subject-by-subject basis. As reported at these meetings one year ago, the fit mean square statistics and their standard errors as well as a review of fit statistics for individual items indicated that the vast majority of items, when calibrated by subject, fit the model. However, conceptually it just does not seem as though a 1,000 item interdisciplinary basic science examination with subjects as diverse as biochemistry and the behavioral sciences should be unidimensional. The same can be said of the interdisciplinary 900 item clinical science examination with subjects ranging from knowledge of surgery to knowledge of psychiatry. While conceptually there would appear to be differences among these disciplines, what is required by the Rasch model is that unidimensionaltiy be demonstrated empirically or statistically.

To test for the unidimensionality of multi-subject content, the item difficulties resulting from the subject calibrations were linked to those from the total interdisciplinary examination calibrations. Items that do not fit the total test dimension are identified by their misfit in the link. Also, statistical indicies of the quality of the link are provided. A high quality

link could occur only if the subject items and other items from the total examination measure the same dimension.

Figure 1 shows the computer printed output of the program LINK used in the manner just described. One axis is for item difficulties from one of the clinical subtest calibrations and the other axis is for item difficulties from the total test calibration. This plot of subject and total test calibrations yields essentially a straight line with no items that are really poor fits. The band shown on the plot is one logit wide, half a logit on either side of the line of perfect fit. This plot is quite similar to those for the other subjects. Five statistical indices of the quality of the link accompany each plot. For each link, all five indices indicate a quality link. We have concluded that the various basic science subjects and the various clinical science subjects as represented on National Board examinations each meet the statistical criteria for unidimensionality as required by the Rasch model.

Now, back to specific equating procedures. We are now in the process of moving the item difficulties of all basic science questions onto a bank scale. Later the same will be done for the clinical science questions. Thus, our use of the Rasch model for equating examinations has passed through a three-phase evolution. First, was the equating of current and reference tests via a combined common person, common item procedure based on subtest calibrations of item difficulties. Next came the equating of current and reference tests via a common item procedure based on total test calibrations of item difficulties. Finally, just ahead, is the common item equating of a current examination, with difficulties calibrated on the total test, to a combination of several reference examinations or, in other words, to the item bank.

This last step, equating to the bank scale, is important because it will allow for maximum flexibility in use of the item bank for test design and development. It will also provide a broader examinee base for the conversion of Rasch person abilities to scores on a traditional scale. Because of the relatively large number of both link items and examinees as well as the statistically documented unidimensionality of the test content, the network of links required to establish the bank scale of item difficulties is not likely to produce links that tend to be incoherent as was the case with some of the links with fewer than 10 items that were discussed by Frank Hughes in the previous paper.

Once the item difficulties are in place on the bank scale, it will not be assumed that they can be used forever without being checked. Procedures to detect drift in item difficulty and fit will be implemented.

Use of the Rasch model in this way will enable us to monitor, over time, comparability of examinee samples and the difficulty of items. This, in turn, will greatly enhance our ability to monitor the comparability over time of the standards set for passing the tests for purposes of certification. It is this continual efficient monitoring of examinee ability, item difficulty, item fit and also person fit, which incidentally has not been addressed by any of these equating papers, that makes the Rasch model so attractive, not only for equating tests but for use in many aspects of an examination program.

COMBINED COMMON PERSON AND COMMON ITEM
EQUATING OF MEDICAL SCIENCE EXAMINATIONS

Paul R. Kelley
National Board of Medical Examiners

TABLE 1

Summary of Equating Shifts Expressed in Logits

Subtest	Reference Test Mean Abilities (Common Persons)	Reference - Current Mean Item Difficulties (Common Items)	Shift Derived from Combined Common Persons and Common Items*
Anatomy	.51	.01	.50
Physiology	.79	.29	.50
Biochemistry	.77	-.04	.81
Pathology	1.08	.18	.90
Microbiology	.68	-.12	.80
Pharmacology	.74	.18	.56
Behavioral Science	.81	-.06	.87

*The shift required in the current item difficulties to equate them to a common origin on the reference examination scale. These figures are equal to column 1 values minus the corresponding column 2 values.

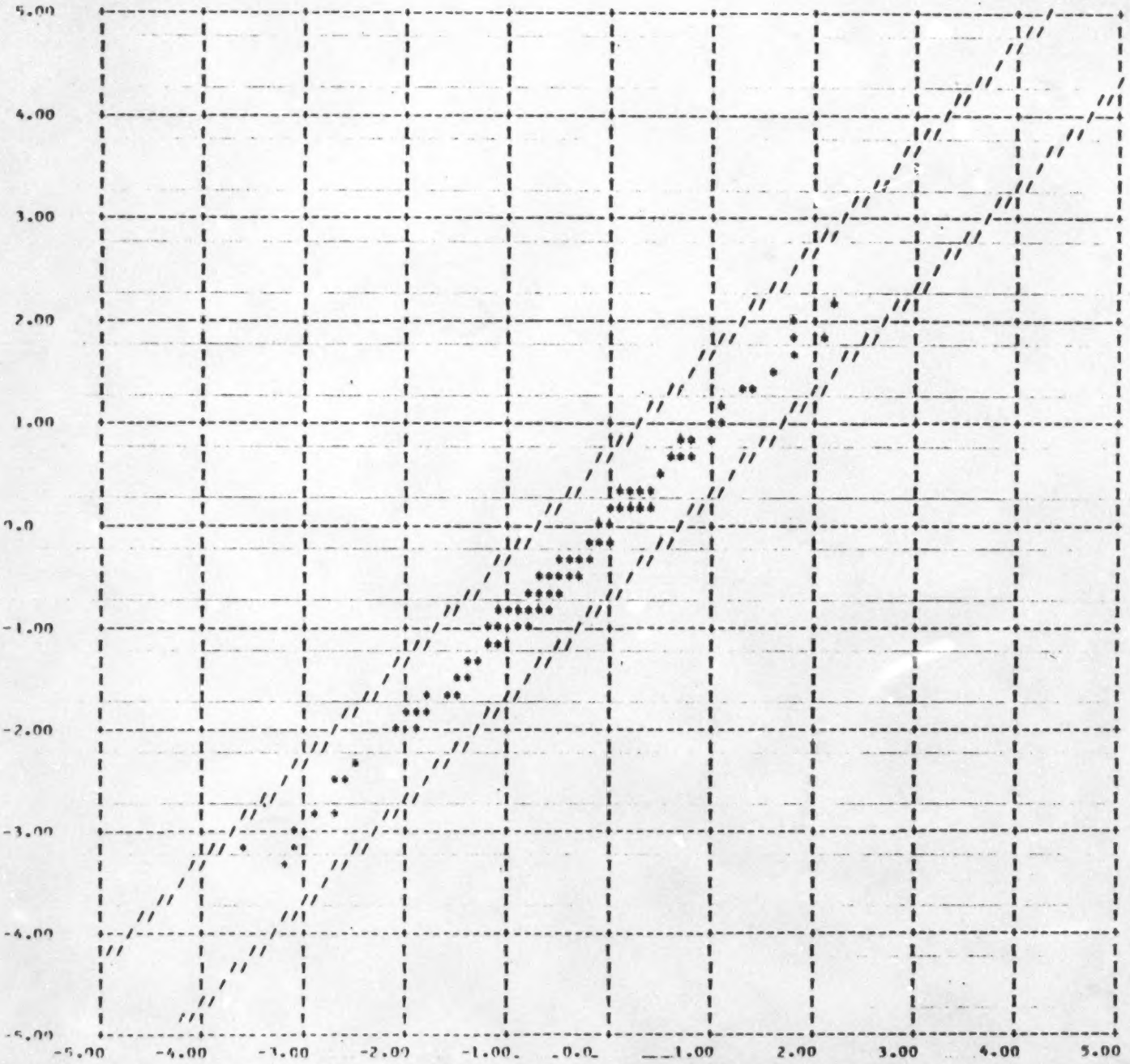

FIGURE 1 - COMBINED COMMON PERSON AND COMMON ITEM EQUATING OF MEDICAL SCIENCE EXAMINATIONS
Paul R. Kelley, National Board of Medical Examiners